# Hierarchical Wavelet Modelling of Environmental Sensor Data

Yann Ruffieux and A. C. Davison*

December 31, 2007

**Summary**

Motivated by the need to smooth and to summarize multiple simultaneous time series arising from networks of environmental monitors, we propose a hierarchical wavelet model for which estimation of hyperparameters can be performed by marginal maximum likelihood. The result is an empirical Bayes thresholding procedure whose results improve on those of wavethresh in terms of mean square error. We apply the approach to data from the SensorScope environmental modelling system, and briefly discuss issues that arise concerning variance estimation in this context.

**Keywords:** Empirical Bayes; Environmental sensor; Hierarchical model; Mixture model; Normal distribution; SensorScope; Spike-and-slab model; Wavelet.

*Institute of Mathematics, IMA-FSB-EPFL, Station 8, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, Anthony.Davison@epfl.ch

# 1 Introduction

Realistic environmental modelling depends critically on detailed data that have hitherto been too expensive to obtain. For example, in temperature or rainfall studies it has not been uncommon to base modelling on time series of daily values gathered many kilometres apart, so that interpolation—so-called downscaling—has been required to make inferences at more local scales, both in time and in space. The stunning decrease in the cost and increase in the quality of telecommunications and other electronic equipment now make it feasible to create dense wireless networks of cheap sensors which yield measurements at high spatial and temporal resolutions, with the potential to learn vastly more about the detailed working of environmental phenomena. SensorScope (`http://sensorscope.epfl.ch`) is an interdisciplinary project at EPFL which provides such data, the quantity and quality of which create new problems for statistics: of design of such networks; of data treatment; and of interpretation. Below we describe an approach to summarization of parallel data from many sensors, preliminary to more detailed exploitation.

Before deployment in more demanding surroundings, in July 2006 a working group embarked on a project to measure numerous atmospheric variables around the campus. The Lausanne Urban Canopy Experiment (LUCE) consisted of a network of around 100 weather stations, deployed within an area of roughly one-half of a square kilometre. Stations with various different configurations were used, collecting data on quantities such as air temperature, ground temperature, soil moisture, humidity, rainfall, and wind speed and direction, in real time and at short time intervals. The resulting data are noteworthy: they are highly localized both in time and space; they run over several months; and they are gathered in an unusual urban-type environment. Similar dense networks have been or are being deployed in more taxing settings, for example at the Grand-St-Bernard pass (altitude 2400m) and at the Plaine Morte glacier (altitude 2750m); more details can be obtained from the link given above.

Figure 1 displays ambient temperatures taken during the LUCE experiment at four stations in a time frame of 24 hours (midnight to midnight); the measurements were made roughly every 30 seconds. The curves show a common trend due to the high proximity of the stations but have distinct features due to their particular surroundings—exposure, altitude, ground type, and so forth. The signals have appreciable local variation and some kind of smoothing seems required.

Although smoothing the signals independently would be quick and easy, we would like to allow for the obvious similarity between signals from different stations. One way to model this is to assume the existence of an underlying curve that generates the smooth station-specific curves. We will concentrate on the air temperature measurements, as a

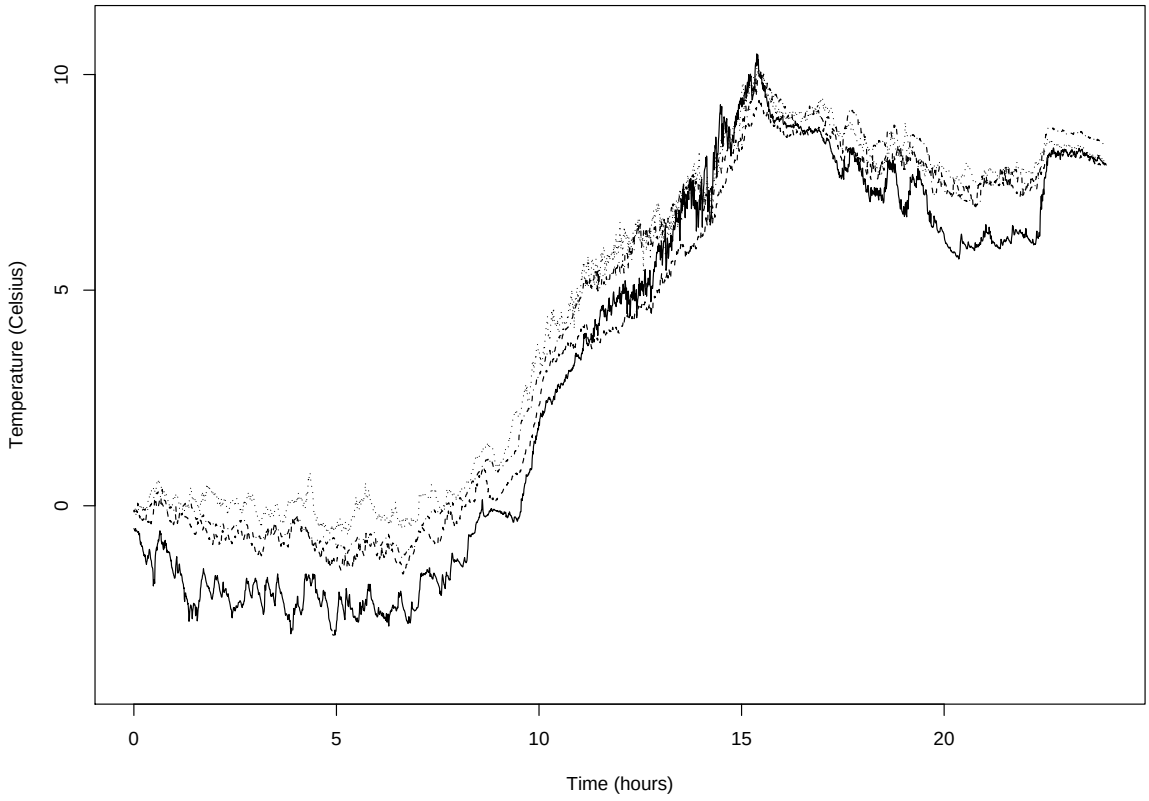


Figure 1: Twenty-four hours of air temperature measurements (°C) at four SensorScope weather stations on the EPFL campus.

prototype for other series of continuous values. We define temperature curves $Y_s(t)$ for each of the $S$ weather stations, and assume that

$$Y_s(t) = \mu(t) + \psi_s(t) + e_s(t), \quad s = 1, \ldots, S.$$

The time variable $t$ can be taken over any interval, but for our purposes it covers 24 hours, from midnight to midnight. The function $\mu(t)$ can be regarded as the underlying temperature curve for the whole campus and $\psi_s(t)$ as the effect from station $s$, while $e_s(t)$ represents the residual effect function at station $s$. Thus the function $\mu(t) + \psi_s(t)$ can be considered to be the noise-free curve for station $s$.

In practice data may be observed at irregular intervals, with the observation times varying from station to station. For later developments we interpolate the observed time series to obtain observations on a regular grid $t_1, \ldots, t_N$ over the whole day, with $t_{i+1} - t_i = c$ for all $i$ and with $N = 2^J$ for a fixed positive integer $J$. The model may

then be written

$$\mathbf{Y}_s \;=\; \boldsymbol{\mu} + \boldsymbol{\psi}_s + \boldsymbol{e}_s, \quad s = 1, \ldots, S, \tag{1}$$

with $\mathbf{Y}_s = [Y_s(t_1), \ldots, Y_s(t_N)]^{\mathrm{T}}$ denoting the vector of observations at station $s$, $\boldsymbol{\mu} = [\mu(t_1), \ldots, \mu(t_N)]^{\mathrm{T}}$, $\boldsymbol{\psi}_s = [\psi_s(t_1), \ldots, \psi_s(t_N)]^{\mathrm{T}}$, and $\boldsymbol{e}_s = [e_s(t_1), \ldots, e_s(t_N)]^{\mathrm{T}}$. Our goal here will be to filter out the noise $\boldsymbol{e}_s$ in order to find smooth estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\psi}_s$.

Equations (1) can be written in matrix form as

$$Y \;=\; XB + E, \tag{2}$$

where $Y$ is the $S \times N$ matrix with the observation vectors $\mathbf{Y}_s$ in rows, $X = [\mathbf{1}_S \mid I_s]$ is a $S \times (S+1)$ design matrix, $B$ is a $(S+1) \times N$ matrix with $\boldsymbol{\mu}$ in the first row and $\boldsymbol{\psi}_s$ in row $s+1$, and $E$ is the $S \times N$ matrix of residuals. Here and elsewhere, the term $\mathbf{1}_S$ designates a $S \times 1$ vector of ones and $I_S$ is the $S \times S$ identity matrix. Note that (2) corresponds to a particular case of the functional mixed-effect model described in Morris and Carroll (2006), but with no random effect.

One of our aims is to develop an automatic approach: ultimately we would like to produce the necessary estimates for SensorScope in real time without any intervention. Many of the assumptions below are made towards that end, perhaps at the expense of generality. In Section we outline our model and then in Section 3 we discuss inference for its hyperparameters. Section 4 describes a small simulation study to compare our approach with use of wavethresh, and is followed by an application to the SensorScope data. The paper concludes with a brief discussion.

# 2 Wavelet regression and Bayesian modelling

We work in the wavelet domain, which gives an attractive basis for curve regularization and modelling. Wavelets have received a lot of attention from the mathematical and statistical communities in the past few years, and much has been written about them. Strang (1993) gives a nice introduction, while Percival and Walden (1993) provide a detailed account focusing on time series analysis.

Let $W$ be the $N \times N$ orthogonal discrete wavelet transform (DWT) matrix under a given wavelet basis. Then right-multiplication of a $N \times 1$ vector by $W^{\mathrm{T}}$ corresponds to a change of basis from the time domain to the wavelet domain. The coefficients in this new basis have a specific interpretation in terms of local variation of the signal, much like Fourier coefficients. However the wavelet basis is more descriptive because it allows the local variation to depend on location, or in our context, time. There is an infinite

number of possible wavelet bases, but only a few of them are widely used (Daubechies, 1992). At this point no assumption is needed on the type of wavelet used, and $W$ can be seen as a generic wavelet change of basis. Wavelet coefficients can be computed in a highly efficient manner using the pyramid algorithm (Mallat, 1989).

Right-multiplying both sides of (2) by $W^{\mathrm{T}}$ yields

$$D \;=\; XB^* + E^*, \tag{3}$$

where $D = YW^{\mathrm{T}}$ is a $S \times N$ matrix whose row $s$ contains the 'observed' wavelet coefficients for station $s$, $B^* = BW^{\mathrm{T}}$ contains the wavelet coefficients for the mean function $\boldsymbol{\mu}$ in the first row and the wavelet coefficients for the station-specific effects $\boldsymbol{\psi}_s$ in the other rows, while $E^* = EW^{\mathrm{T}}$ contains the residuals in the wavelet space. We double-index the rows of $D, B^*$, and $E^*$ to include the scale and location of each coefficient: for $j = 0, 1, \ldots, J-1$, $k = 1, 2, \ldots, 2^j$ and $s = 1, 2, \ldots, S$, write $d_{jk}^{(s)}$ as the observed coefficient at scale $j$ and location $k$ in the wavelet decomposition of $\mathbf{Y}_s$. Then a coefficient with large $j$ has high resolution and so is more likely to be interpreted as noise. Similarly, define $\theta_{jk}^{(s)}$ and $\varepsilon_{jk}^{(s)}$ as the $(j,k)$-coefficient in the wavelet decompositions of $\boldsymbol{\psi}_s$ and $\boldsymbol{e}_s$ respectively. Finally set $\zeta_{jk}$ as the $(j,k)$-coefficient in the decomposition of the mean vector $\boldsymbol{\mu}$. For fixed $j, k$, and $s$, the relation between these coefficients is

$$d_{jk}^{(s)} \;=\; \zeta_{jk} + \theta_{jk}^{(s)} + \varepsilon_{jk}^{(s)}. \tag{4}$$

The $\varepsilon_{jk}^{(s)}$ are assumed to be mutually independent, identically distributed variables from a centred Gaussian density of variance $\sigma^2$. This follows from our assumptions on the time-domain residuals $\boldsymbol{e}_s$.

We now set priors on the parameters $\zeta_{jk}$ and $\theta_{jk}^{(s)}$. We first define a Bernoulli random variable $Z_{jk}$, with $P(Z_{jk} = 1) = \pi_j$ and $P(Z_{jk} = 0) = 1 - \pi_j$, and link it to the coefficients as follows:

- if $Z_{jk} = 1$, then we independently set $\zeta_{jk} \sim \mathcal{N}(0, \tau_j^2)$ and $\theta_{jk}^{(s)} \sim \mathcal{N}(0, \eta^2)$ for $s = 1, \ldots S$;
- if $Z_{jk} = 0$, then $\zeta_{jk}$ and each $\theta_{jk}^{(s)}$ follow a degenerate distribution with unit mass at zero.

This latent variable $Z_{jk}$ thus indicates whether the corresponding coefficients $\zeta_{jk}$ and $\left\{\theta_{jk}^{(s)}, s = 1, \ldots S\right\}$ are 'switched on'. We use microarray terminology and say that the $(j,k)$-coefficients are *differentially expressed* if $Z_{jk} = 1$. The main purpose in making the above assumptions is that we consider that only those coefficients which are differentially expressed contain information on the smooth signal, the rest being treated as noise. In

particular, we expect $\pi_j$ to decrease with $j$ so that high-resolution coefficients are less likely to be differentially expressed *a priori*.

On marginalizing over $Z_{jk}$ we see that the coefficients follow mixture distributions, that is,

$$\zeta_{jk} \sim \pi_j \mathcal{N}(0, \tau_j^2) + (1 - \pi_j)\delta_0$$

and

$$\theta_{jk}^{(s)} \sim \pi_j \mathcal{N}(0, \eta^2) + (1 - \pi_j)\delta_0,$$

with $\delta_0$ denoting the distribution with unit mass at zero. These prior assumptions are very similar to those of Morris and Carroll (2006), who also set a mixture model on the elements of $B$. An important difference here is that for fixed $(j, k)$, the coefficients $\zeta_{jk}$ and $\theta_{jk}^{(s)}, s = 1, \ldots, S$ are marginally dependent: *a priori* they are either all differentially expressed or all zero. This mixture approach was also used by Abramovich *et al.* (1998) and Johnstone and Silverman (2005), but applied to single times series.

For completeness, we place a vague prior on the single scaling coefficient of each wavelet decomposition. Thus these are estimated *a posteriori* by the sample scaling coefficients, typically the sample mean multiplied by the square root of the number of coefficients.

Now we establish the posterior distribution of the wavelet coefficients given the observed coefficients $D$ and the hyperparameters $\sigma^2, \eta^2, \tau_j^2$, and $\pi_j$. Set $\mathbf{d}_{jk} = \left[d_{jk}^{(1)}, d_{jk}^{(2)}, \ldots, d_{jk}^{(S)}\right]^{\mathrm{T}}$ (a column of the matrix $D$) and $\boldsymbol{\theta}_{jk} = \left[\theta_{jk}^{(1)}, \theta_{jk}^{(2)}, \ldots, \theta_{jk}^{(S)}\right]^{\mathrm{T}}$. It is easily seen that, given $Z_{jk} = 1$, the joint distribution of $\left[\mathbf{d}_{jk}^{\mathrm{T}}, \boldsymbol{\theta}_{jk}^{\mathrm{T}}, \zeta_{jk}\right]^{\mathrm{T}}$ is a centred multivariate Gaussian distribution, and that it is degenerate given $Z_{jk} = 0$. In particular, the marginal density function of $\mathbf{d}_{jk}$ is the mixture

$$p(\mathbf{d}_{jk}) \;=\; \pi_j g(\mathbf{d}_{jk}; 0, A_j) + (1 - \pi_j) g(\mathbf{d}_{jk}; 0, \sigma^2 I_S), \tag{5}$$

where $g(\,\cdot\,; \boldsymbol{m}, \Sigma)$ is the density function of a multivariate Gaussian random variable with mean $\boldsymbol{m}$ and covariance matrix $\Sigma$, and where

$$A_j = (\sigma^2 + \eta^2) I_S + \tau_j^2 \mathbf{1}_S \mathbf{1}_S^{\mathrm{T}}$$

is an equicorrelation matrix, whose inverse and determinant can be computed directly using a closed form. It is straightforward to find the posterior odds $\omega_{jk}$ of the $(j, k)$-coefficients being differentially expressed:

$$\omega_{jk} \;=\; \frac{P(Z_{jk} = 0 \mid \mathbf{d}_{jk})}{P(Z_{jk} = 1 \mid \mathbf{d}_{jk})} = \frac{1 - \pi_j}{\pi_j} \times \frac{g(\mathbf{d}_{jk}; 0, \sigma^2 I_S)}{g(\mathbf{d}_{jk}; 0, A_j)}.$$

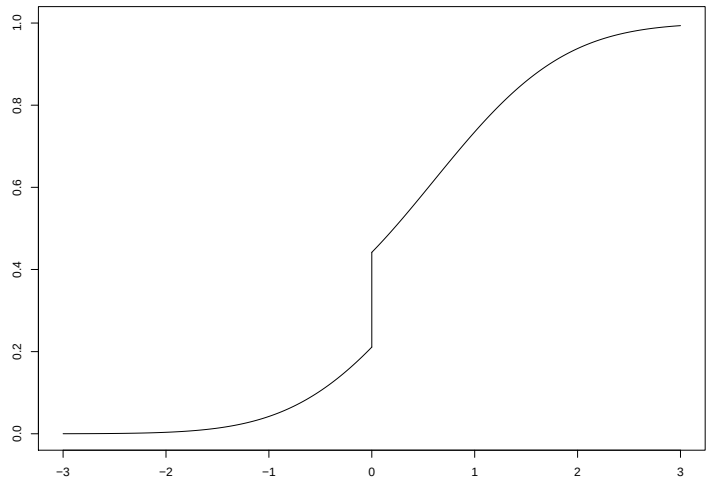


Figure 2: The distribution function of a mixture between a normal distribution and a point mass at zero.

Using properties of the multivariate normal distribution, we find that the posterior distribution function of $\zeta_{jk}$ is the mixture

$$F(\zeta_{jk} \mid \mathbf{d}_{jk}) \;=\; \frac{1}{1+\omega_{jk}}\Phi\left\{\frac{\zeta_{jk}-\tau_j^2\mathbf{1}_S^{\mathrm{T}}A_j^{-1}\mathbf{d}_{jk}}{\tau_j\sqrt{1-\tau_j^2\mathbf{1}_S^{\mathrm{T}}A_j^{-1}\mathbf{1}_S}}\right\}+\frac{\omega_{jk}}{1+\omega_{jk}}I\{\zeta_{jk}\geq 0\}, \qquad (6)$$

where $\Phi$ is the standard normal distribution function and $I$ is the indicator function. Recall that for fixed $s$, the $\theta_{jk}^{(s)}$ represent the coefficients for the *effect* of station $s$. We will focus on the estimation of $\nu_{jk}^{(s)} = \zeta_{jk}+\theta_{jk}^{(s)}$: for fixed $s$, the $\nu_{jk}^{(s)}$ are the wavelet coefficients for the noise-free signal from station $s$. Let $\mathbf{c}_{js} = \left[\tau_j^2,\ldots,\tau_j^2,\tau_j^2+\eta^2,\tau_j^2,\ldots,\tau_j^2\right]^{\mathrm{T}}$ be a $S\times 1$ vector, with the $\tau_j^2+\eta^2$ term in position $s$. The posterior distribution of $\nu_{jk}^{(s)}$ is

$$F\left(\nu_{jk}^{(s)} \mid \mathbf{d}_{jk}\right) \;=\; \frac{1}{1+\omega_{jk}}\Phi\left\{\frac{\nu_{jk}^{(s)}-\mathbf{c}_{js}^{\mathrm{T}}A_j^{-1}\mathbf{d}_{jk}}{\sqrt{\eta^2+\tau_j^2-\mathbf{c}_{js}^{\mathrm{T}}A_j^{-1}\mathbf{c}_{js}}}\right\}+\frac{\omega_{jk}}{1+\omega_{jk}}I\left\{\nu_{jk}^{(s)}\geq 0\right\}. \qquad (7)$$

These distribution functions have jumps of sizes $\omega_{jk}/(1+\omega_{jk})$ at zero; see Figure 2.

# 3 Hyperparameter estimation and inference

## 3.1 Basic model

We use an empirical Bayes approach to choose the hyperparameters $\sigma^2, \eta^2, \tau_j^2$, and $\pi_j$ of our model. The full marginal log likelihood for these hyperparameters is

$$\ell\left(\sigma^2, \eta^2, \tau_0^2, \ldots, \tau_{J-1}^2, \pi_0, \ldots, \pi_{J-1}; D\right) \;=\; \sum_{j,k} \log p(\mathbf{d}_{jk}), \tag{8}$$

where the $p(\mathbf{d}_{jk})$ are computed from (5). The hyperparameter estimates will be the values which maximize (8). The number of hyperparameters, $2(J+1)$, can be relatively high, since typically $J \geq 10$ in this context. We can ease the computation somewhat by following the model of Abramovich *et al.* (1998) for the variance of $\zeta_{jk}$,

$$\tau_j^2 = C \cdot 2^{-j\alpha}, \quad j = 0, 1, \ldots, J-1,$$

for $\alpha,\ C > 0$. Making the appropriate substitutions in (8), we can then maximize over the parameters $\alpha$ and $C$ rather than over $\tau_0^2, \ldots, \tau_{J-1}^2$. Experiments with unconnected $\tau_j^2$ suggest this assumption is reasonable. As we expect only the higher-resolution coefficients to correspond to noise, we can group the $k$ lower scale $\pi_j$'s into a single parameter $\pi_B$, which we expect to be close to unity.

To estimate the wavelet coefficients $\zeta_{jk}$ and $\nu_{jk}^{(s)}$, we follow the Abramovich *et al.* (1998) thresholding approach, by computing the posterior medians of the distributions (6) and (7). We can find the median for the general mixture distribution

$$H(x) = \frac{1}{1+\omega}\Phi\left(\frac{x-\mu}{\nu}\right) + \frac{\omega}{1+\omega} I(x \geq 0)$$

as follows (see Figure 2):

1. if $\omega \geq 1$, then the jump at 0 is greater than $1/2$ and the median is zero;
2. if $\frac{1-\omega}{2} \leq \Phi(-\mu/\nu) \leq \frac{1+w}{2}$, the the jump at 0 starts below $1/2$ and lands above $1/2$, so the median is zero;
3. otherwise the median is

$$\nu\Phi^{-1}\left[\frac{1}{2} + \frac{\omega}{2}\text{sign}\left\{\frac{\Phi(-\mu/\nu)}{1+\omega} - \frac{1}{2}\right\}\right] + \mu.$$

Let $\bar{d}_{jk} = \sum_{s=1}^{S} d_{jk}^{(s)}/S$ be the mean of the $(j,k)$-coefficients across the sites. After

some algebra we find that

$$\hat{\zeta}_{jk} = \text{med}\left(\zeta_{jk} \mid \mathbf{d}_{jk}\right) = \text{sign}\left(\bar{d}_{jk}\right) \max(0, \gamma_{jk}),$$

with

$$\gamma_{jk} = \frac{S\tau_j^2}{\sigma^2 + \eta^2 + S\tau_j^2} \left|\bar{d}_{jk}\right| - \sqrt{\frac{\tau_j^2(\sigma^2 + \eta^2)}{\sigma^2 + \eta^2 + S\tau_j^2}} \times \Phi^{-1}\left\{\frac{1 + \min(\omega_{jk}, 1)}{2}\right\}.$$

The 'threshold' rule is thus

$$\hat{\zeta}_{jk} = 0 \Longleftrightarrow \left|\bar{d}_{jk}\right| \leq \frac{1}{S}\sqrt{\frac{(\sigma^2 + \eta^2 + S\tau_j^2)(\sigma^2 + \eta^2)}{\tau_j^2}} \times \Phi^{-1}\left\{\frac{1 + \min(\omega_{jk}, 1)}{2}\right\}.$$

As $\omega_{jk}$ depends both on $\bar{d}_{jk}$ and on $\sum_{s=1}^{S}\left(d_{jk}^{(s)}\right)^2$, this is not a pure threshold rule.

As a result of taking the posterior median, the estimated coefficients are either differentially expressed or set to zero, as in the prior model. By setting certain coefficients to zero we are removing the noise and retaining just the smooth components of the curves.

Once we have our estimates $\hat{\zeta}_{jk}$ and $\hat{\nu}_{jk}^{(s)}$, we reconstruct the smooth signal by applying the inverse DWT transform to the appropriate vectors of estimated coefficients: the smooth estimated curves $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\psi}}_s$ can be extracted from the matrix $W\hat{B}^*$, where $\hat{B}^*$ is the matrix $B^*$ with the estimates $\hat{\zeta}_{jk}$ and $\hat{\theta}_{jk}^{(s)} = \hat{\zeta}_{jk} - \hat{\nu}_{jk}^{(s)}$ inserted at the appropriate locations.

## 3.2 Non-constant noise variance

Above we assumed a constant variance for the noise. However examination of the data in both time and wavelet domains suggests that there are two variance regimes: one towards the afternoon during which the variance is clearly larger, and another with less variability. The boundaries between them are hard to pinpoint, and vary from day to day. We deal with this by assigning mixtures of normal densities to the errors $\varepsilon_{jk}^{(s)}$, namely:

$$\varepsilon_{jk}^{(s)} \sim \tilde{\pi}\mathcal{N}\left(0, \sigma_1^2\right) + (1 - \tilde{\pi})\mathcal{N}\left(0, \sigma_2^2\right),$$

with $0 < \tilde{\pi} < 1$. We have implicitly defined a Bernoulli variable $\tilde{Z}_{jk}$ such that $\tilde{Z}_{jk} = 1$ implies that the error variance equals $\sigma_1^2$ while $\tilde{Z}_{jk} = 0$ implies it equals $\sigma_2^2$. As in the constant-variance case, $\sigma_1^2$ and $\sigma_2^2$ can be estimated by maximizing the marginal likelihood

$$\ell\left(\sigma_1^2, \sigma_2^2, \eta^2, \tau_j^2, \pi_j, \tilde{\pi}; D\right) = \sum_{j,k} \log p(\mathbf{d}_{jk}),$$

with

$$\begin{aligned} p(\mathbf{d}_{jk}) &= \pi_j\tilde{\pi}g(\mathbf{d}_{jk};0,A_{1j}) + \pi_j(1-\tilde{\pi})g(\mathbf{d}_{jk};0,A_{2j}) \\ &\quad +(1-\pi_j)\tilde{\pi}g(\mathbf{d}_{jk};0,\sigma_1^2 I_S) + (1-\pi_j)(1-\tilde{\pi})g(\mathbf{d}_{jk};0,\sigma_2^2 I_S), \end{aligned}$$

where $A_{1j}$ is the matrix $A_j$ with the $\sigma^2$ terms replaced by $\sigma_1^2$, and similarly for $A_{2j}$ and $\sigma_2^2$. The posteriors for $\zeta_{jk}$ and $\nu_{jk}^{(s)}$ are now mixtures of two distinct normal components and a point mass at zero, making computation of the posterior medians awkward. We circumvent this by first computing the posterior odds of the error variance being $\sigma_1^2$, for each $(j,k)$, giving

$$\tilde{\omega}_{jk} = \frac{1-\tilde{\pi}}{\tilde{\pi}} \times \frac{\pi_j g(\mathbf{d}_{jk};0,A_{2j}) + (1-\pi_j)g(\mathbf{d}_{jk};0,\sigma_2^2 I_S)}{\pi_j g(\mathbf{d}_{jk};0,A_{1j}) + (1-\pi_j)g(\mathbf{d}_{jk};0,\sigma_1^2 I_S)}.$$

The estimators are then

$$\hat{\zeta}_{jk} = \text{med}\left(\zeta_{jk} \mid \tilde{Z}_{jk}=1\right) I\left\{\tilde{\omega}_{jk} \le 1\right\} + \text{med}\left(\zeta_{jk} \mid \tilde{Z}_{jk}=0\right) I\left\{\tilde{\omega}_{jk} > 1\right\}$$

and

$$\hat{\nu}_{jk}^{(s)} = \text{med}\left(\nu_{jk}^{(s)} \mid \tilde{Z}_{jk}=1\right) I\left\{\tilde{\omega}_{jk} \le 1\right\} + \text{med}\left(\nu_{jk}^{(s)} \mid \tilde{Z}_{jk}=0\right) I\left\{\tilde{\omega}_{jk} > 1\right\}.$$

This dual-variance model need not be applied at all resolutions, but only at the few highest which directly correspond to noise. One can use the single-variance model (5) for the lower-scale levels.

# 4 Simulation study

Here we apply our estimation approach to simulated data and compare the results with independent wavelet thresholding.

We created $S = 8$ sets of $2^J$ wavelet coefficients following the model described in Section 3, with $J = 11$. More precisely, we generated 'true' coefficients $\zeta_{jk}$ and $\nu_{jk}^{(s)}$ from the appropriate mixture models and added Gaussian noise to obtain the $d_{jk}^{(s)}$. The same approximate coefficient scale was used as that of the coefficients from the SensorScope temperature measurements. Setting $\sigma^2 = 4$, $\eta^2 = 15$, $C = 1000$, and $\alpha = 2$, with $\pi_j$ quickly decreasing in $j$, the reconstructed simulated time series, while very noisy, have a similar underlying smooth curve; three are shown in Figure 3. Applying empirical Bayes estimation yields $\hat{\sigma}^2 = 3.935$, $\hat{\eta}^2 = 12.15$, $\hat{C} = 1162$, and $\hat{\alpha} = 2.729$. The fairly large error in the latter three estimates can probably be explained by the fact that they are based

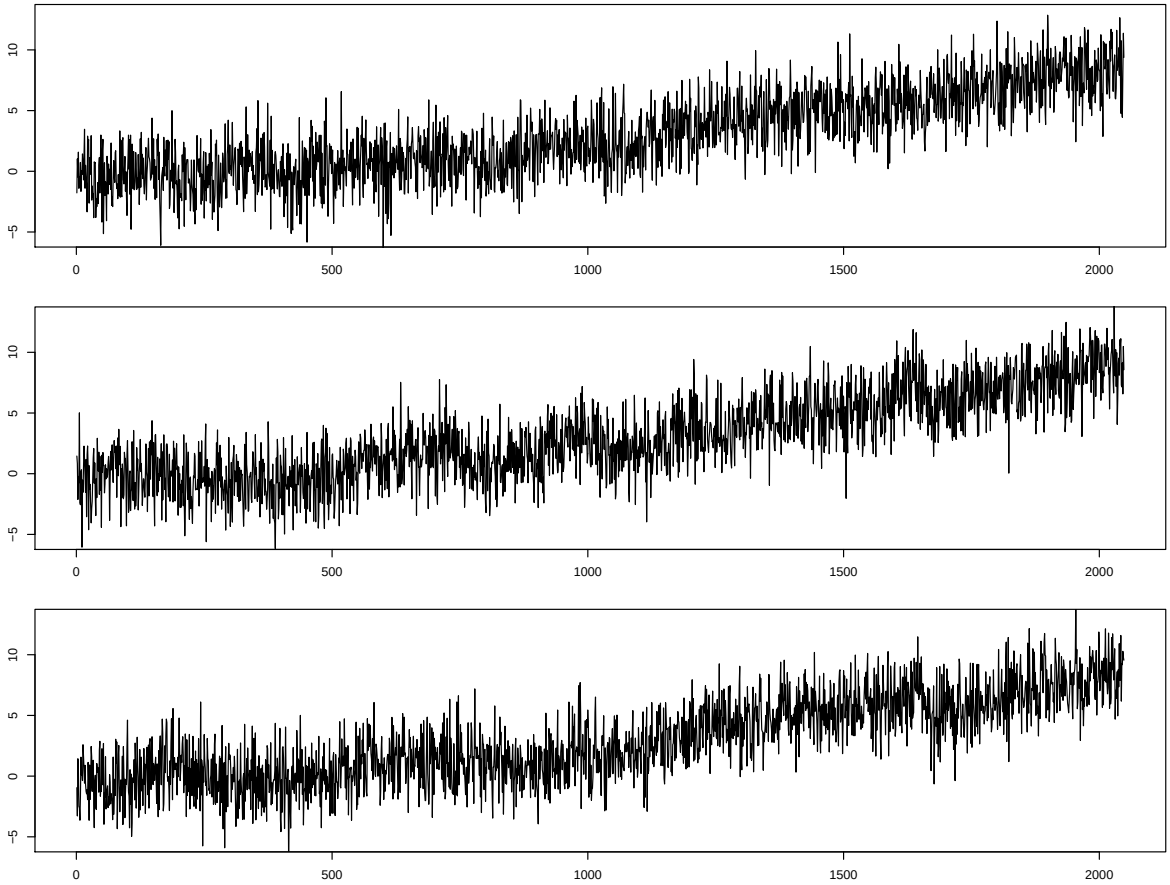


Figure 3: Three time series obtained by applying the inverse DWT to sets of simulated wavelet coefficients.

only on a few low-scale, differentially expressed coefficients, whereas the noise variance estimate is based on numerous high-resolution coefficients. Further simulations based on the above 'true' hyperparameter values suggest that, apart from $\hat{\eta}^2$, the estimates are unbiased. With the values from which we simulate, the empirical Bayes approach seems to underestimate $\eta^2$.

Figure 4 compares the 'true' curves underlying the data in Figure 3, obtained by applying the inverse DWT to the simulated coefficients $\zeta_{jk}$ and $\nu_{jk}^{(s)}$, with the corresponding estimates reconstructed from the posterior medians of (6) and (7), using the empirical Bayes hyperparameter values. None of the estimates $\hat{\nu}_{jk}^{(s)}$ is falsely declared to be differentially expressed or falsely declared to be zero. Of the $2^{11}$ estimates for $\zeta_{jk}$, six are falsely declared to be zero, and one is falsely declared to be differentially expressed. Also shown are the curves smoothed independently using universal 'hard' wavelet thresholding (Nason and Silverman, 1994); this clearly oversmooths, and thereby fails to capture most of the true features of the curves.

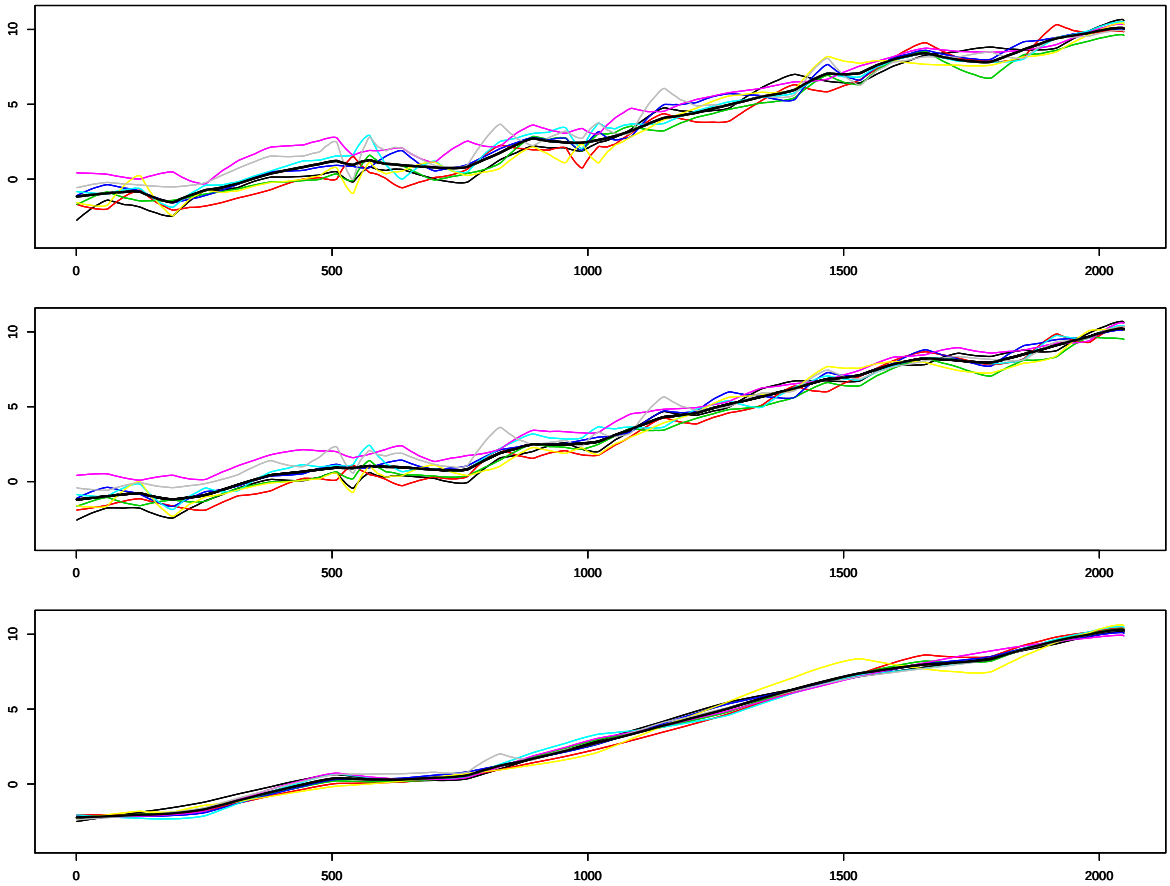


Figure 4: Reconstructed time series obtained by applying inverse DWT to, from top to bottom, the true wavelet coefficients, the coefficients estimated with the median rule via empirical Bayes, and the coefficients after thresholding with wavethresh using default settings. Narrow lines: the eight station-specific smooth signals, thick line: the underlying mean signal (in the bottom panel, the pointwise mean of the station-specific smooth signals).

Given the nature of our model, we expect that a coefficient estimated from a given station may 'borrow strength' from the estimates for the other stations. Thus having data from more stations should give better estimates, both of the underlying coefficients $\zeta_{jk}$ and the station-specific coefficients $\nu_{jk}^{(s)}$. Figure 5 shows the corresponding mean square errors

$$\mathrm{MSE}_\zeta(S) = \frac{1}{2^J} \sum_{j,k} \left( \zeta_{jk} - \hat{\zeta}_{jk,S} \right)^2, \quad \mathrm{MSE}_\nu(S) = \frac{1}{2^J S} \sum_{j,k} \sum_{s=1}^{S} \left( \nu_{jk}^{(s)} - \hat{\nu}_{jk,S}^{(s)} \right)^2,$$

as functions of the number of stations $S$, where $\hat{\zeta}_{jk,S}$ and $\hat{\nu}_{jk,S}^{(s)}$ are the estimates based on $S$ signals. As we increase $S$, we do not re-simulate all the previous coefficients,

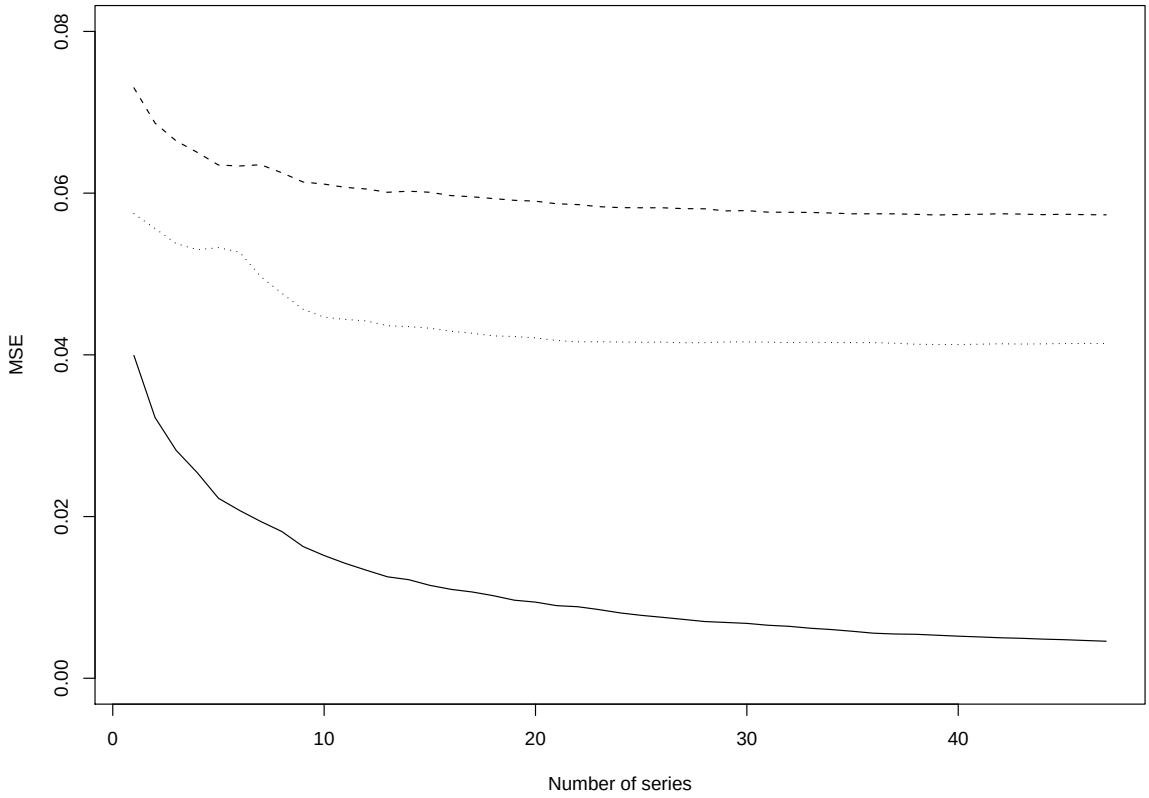


Figure 5: Mean square errors (MSE) for different estimation methods, based on 50 replicates with the number of stations varying from 4 to 50. The lines show: MSE of the posterior median estimates (solid) and wavethresh estimates (dashes) of the grand mean coefficients $\zeta_{jk}$, with the wavethresh estimates simply the mean of the thresholded coefficients; and MSE of the post median estimates of the signal-specific coefficients $\nu_{jk}^{(s)}$ (dots). wavethresh estimates for these have an MSE of about 0.14, which does not decrease with the number of stations.

but simply augment the data by adding further series of coefficients. The values for the hyperparameters were the same as above. We see that $\mathrm{MSE}_\zeta(S)$ decreases roughly exponentially. The evolution of the signal-specific estimates is less clear: the 'borrowing of strength' seems to kick in at around $S = 12$ before stabilizing. We would not see these improvements in performance if the signals were modelled independently.

# 5 Application to SensorScope measurements

We now consider several ways of smoothing the SensorScope time series. The different approaches were applied to all $S = 73$ series for which data were available that day,

though for clarity only eight will be plotted. We use Daubechie's (1992) least-asymmetric compactly supported wavelet with four vanishing moments.

Figure 6 shows the eight time series and a wavelet decomposition of one of them; note the diurnal fluctuations in the coefficient sizes at the lowest level. Figure 7 shows the result of applying the mixed variance model of Section 3.2 to the two finest coefficient levels. There is some smoothing of the individual original series, but little visual difference between this approach and the use of a single variance. The global curve estimate appears slightly biased because it is based on all 73 series, not merely on those plotted. The effect of the mixture is clearer in the wavelet domain: without the mixture, some coefficients at levels 9 and 10 survive thresholding, but with the mixture they are all zeroed out. As expected, the empirical Bayes estimates for the $\pi_j$ decrease rapidly with the level $j$. We have merged the $\pi_j$ for the 5 coarsest levels into a single parameter, whose estimate is practically unity; the estimates for $\pi_9$ and $\pi_{10}$ are both very close to zero. Moreover we find $\hat{\eta}^2 = 0.98$, $\hat{C} = 41761$, and $\hat{\alpha} = 3.15$. As for the variance mixture parameters, the estimates of $\sigma_a^2$, $\sigma_b^2$, and $\tilde{\pi}$ are 0.0036, 0.028, and 0.75 respectively.

The top panel of Figure 8 displays the wavethresh smoothing of the eight time series using the default 'hard' thresholding and error variance estimate based on the squared median absolute deviation of the coefficients at level 3 *and higher*. The results seem unsatisfactory in comparison with those of the hierarchical model: more detailed noise has survived thresholding. The middle panel shows the results of wavethresh when the error variance estimate is based on variance of the coefficients. The choice of variance estimate is crucial: with this much larger variance estimate the curves are appreciably smoother, with so few coefficients surviving that some individual wavelets can be identified. For some purposes these smoother curves might be preferred. The lower panel of the figure shows the results of our hierarchical approach with the noise variance fixed as the variance of all coefficients at levels 3 and higher; though not so smooth as in the second panel, the result might for some purposes be regarded as more satisfactory than the upper panel of Figure 7.

# 6 Discussion

We have described a hierarchical model for treating numerous mutually dependent time series. An empirical Bayes approach was used to assign values to the hyperparameters of the model, and the wavelet coefficients were estimated using a posterior median rule.

Experiments with simulated data show a clear gain in treating the time series simultaneously rather than independently. Any gain is less obvious when applying our

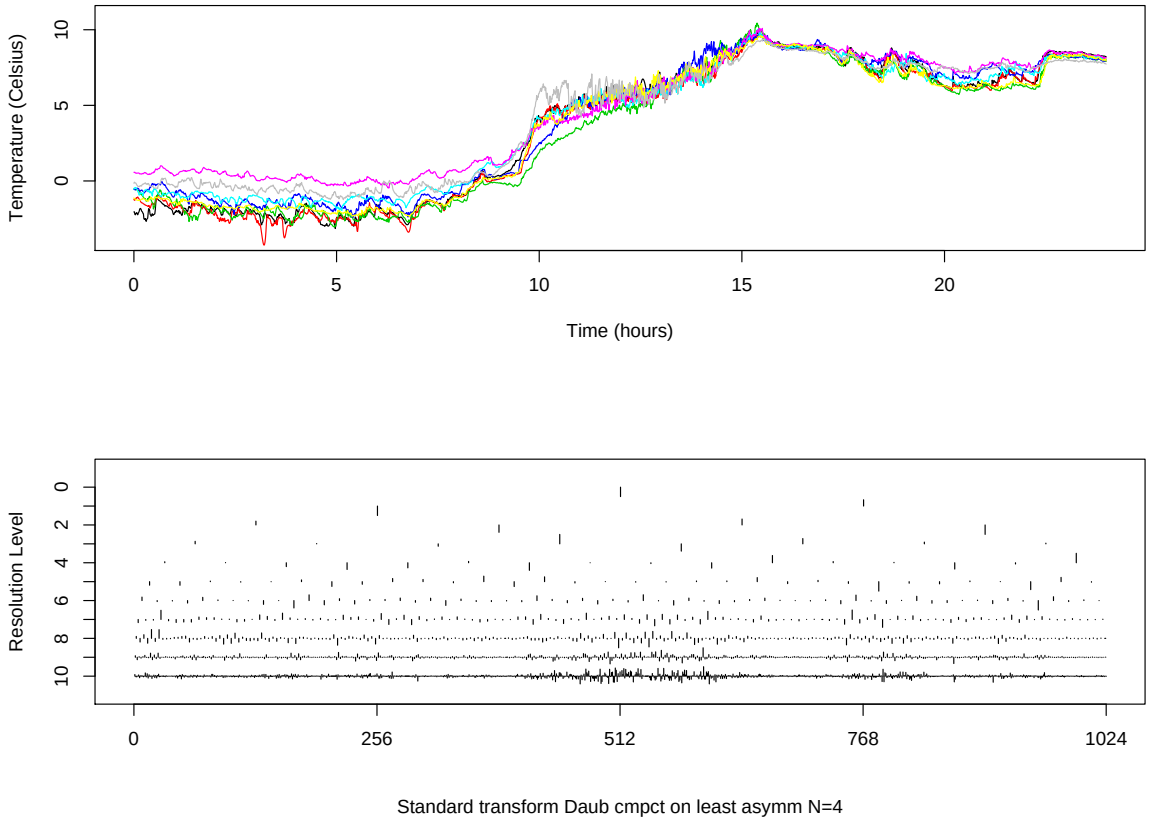


Figure 6: Example series from the SensorScope data. Top: eight time series consisting of SensorScope air temperature data taken over 24 hours. Bottom: wavelet decomposition of one of the series.

methodology to the SensorScope data. The empirical Bayes estimate for the error variance is relatively small, so many high level coefficients pass the thresholding filter. A potential solution would be to assign a value to the error variance beforehand—for example the variance of all the coefficients at level $j_0$ and higher. In this case $j_0$ might be interpreted as a smoothing parameter, a 'slider' that determines the strength of the smoothing.

One natural extension is to allow for a different variance at each station in the basic model (1). To test the effect of this we estimated the variances for each series using the median absolute deviation of its finest level of wavelet coefficients, rescaled the wavelet coefficients to have unit standard error, and applied the mixture model approach to the result, finally back-transforming to allow for the different variances. This had little effect on the reconstructed curves, however.

Other possibilities for work on the rich SensorScope database are the inclusion of covariates to allow for the particular surroundings of each station, and the treatment of

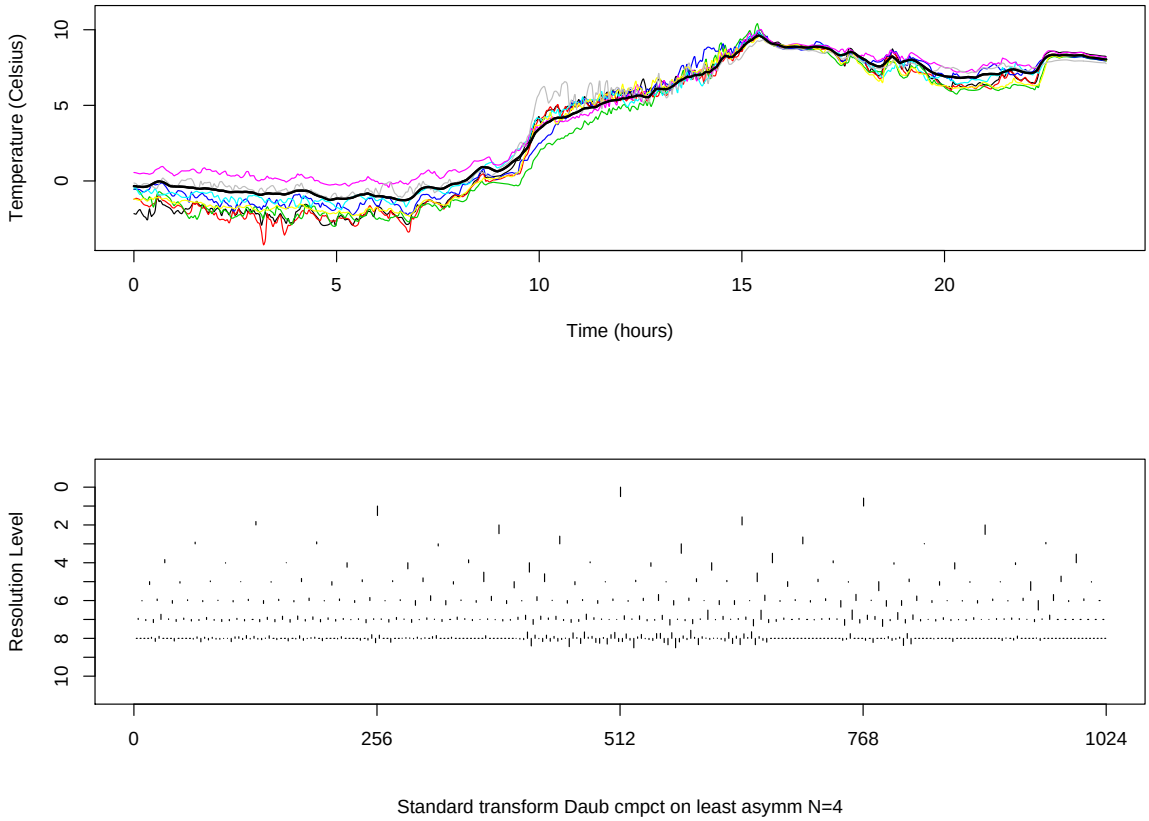


Figure 7: Application of the hierarchical model to the SensorScope data. Top: reconstruction after applying the hierarchical model with a mixture for the error variance; the dark line represents the global curve; bottom: corresponding estimates of the wavelet coefficients $\zeta_{jk}$.

outliers and missing data.

## Acknowledgements

We thank the Swiss National Science Foundation and the ETH domain Centre for Competence in Environment and Sustainability EXTREMES for financial support.

# References

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998) Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society series B* **60**, 725–749.

Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

Johnstone, I. M. and Silverman, B. W. (2005) Empirical Bayes selection of wavelet thresholds. *Annals of Statistics* **33**, 1700–52.

Mallat, S. G. (1989) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

Morris, J. S. and Carroll, R. J. (2006) Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* **68**, 179–199.

Nason, G. P. and Silverman, B. W. (1994) The discrete wavelet transform in `S`. *Journal of Computational and Graphical Statistics* **3**, 162–191.

Percival, D. B. and Walden, A. T. (1993) *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge: Cambridge University Press.

Strang, G. (1993) Wavelet transforms versus Fourier transforms. *Bulletin of the American Mathematical Society* **28**, 288–305.

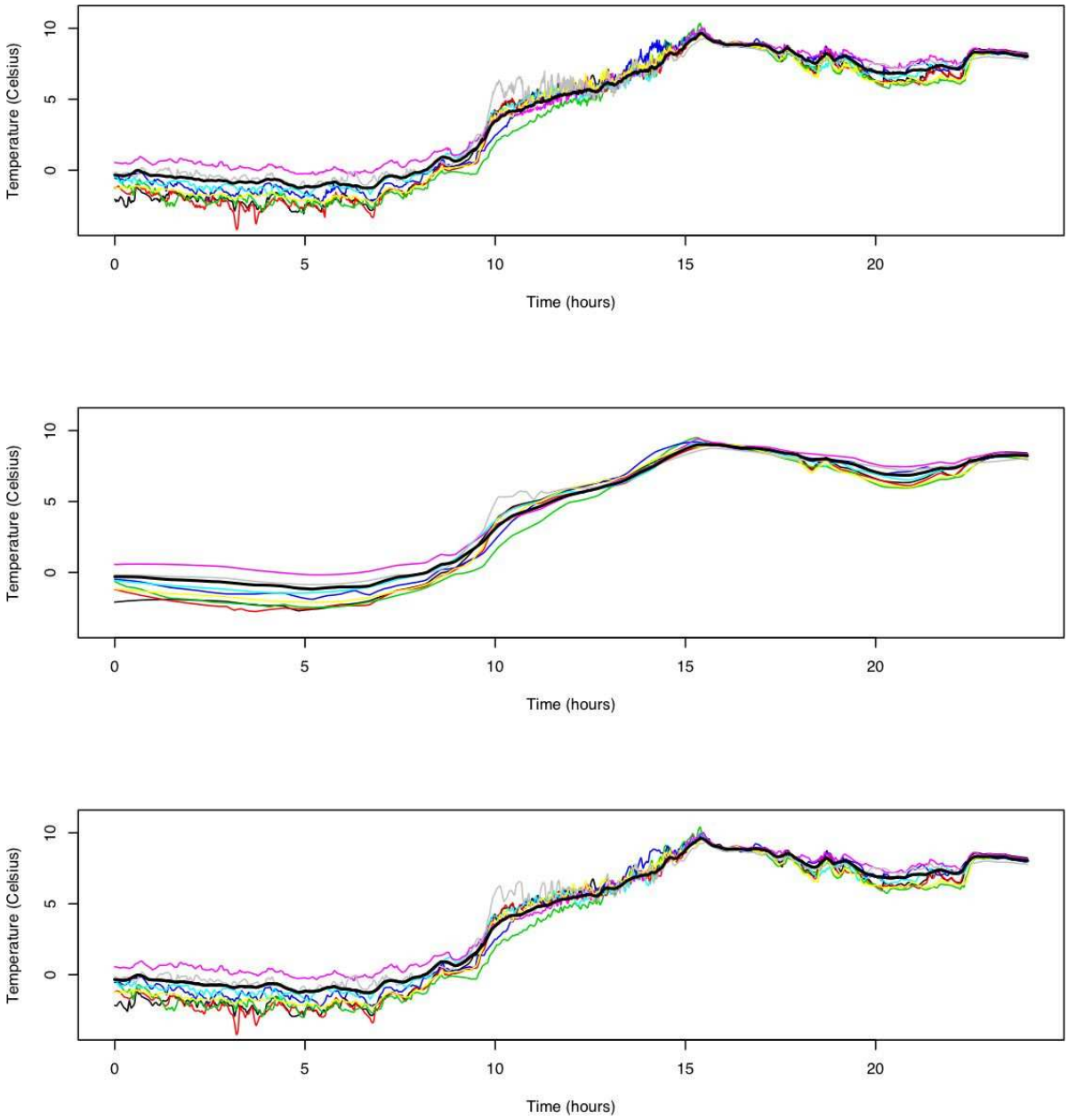


Figure 8: Effect of variance estimation on reconstruction. Top: wavethresh smoothing using default settings; middle: wavethresh smoothing with modified settings; bottom: reconstruction after applying the hierarchical model, with the error variance fixed beforehand as the variance of the coefficients at level 3 and higher.